The emergence of an abstract grammatical category in children's early speech

Stephan C. Meylan

Department of Psychology, University of California, Berkeley

Michael C. Frank

Department of Psychology, Stanford University

Brandon C. Roy

Media Lab, Massachusetts Institute of Technology

Roger Levy

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Thanks to Charles Yang for discussion of his model and data preparation, to Steven Piantadosi for initial discussions, and to the members of the Language and Cognition Lab at Stanford and the Computational Cognitive Science Lab at U.C. Berkeley for valuable feedback. This material is based upon work supported by a National Science Foundation Graduate Research Fellowship to S.C.M. under Grant No. DGE-1106400. R.L. also acknowledges support from Alfred P. Sloan Research Fellowship FG-BR2012-30 and from a fellowship at the Center for Advanced Study in the Behavioral Sciences.

Address all correspondence to Stephan C. Meylan. E-mail: `smeylan@berkeley.edu`.

Roy, B., Frank, M., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663-12668.

Roy, B., Frank, M., & Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. In *Proceedings of the 31st annual meeting of the cognitive science society.*

Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent–child discourse. In *Children's Language* (Vol. 4). Lawrence Erlbaum Associates.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of childes transcripts. *Journal of Child Language*, *37*(3), 705–729.

Suppes, P. (1974). The semantics of children's language. *American Psychologist*, *29*, 103-114.

Theakston, A., Lieven, E., Pine, J., & Rowland, C. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, *28*, 127–152.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL* (p. 252-259).

Vosoughi, S., & Roy, D. (2012). An automatic child-directed speech detector for the study of child language development. In *Proceedings of Interspeech.*

Yang, C. (2013). Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*.

Abstract

How do children begin to use language to say things they have never heard before? The origins of linguistic productivity have been a subject of heated debate: While generativist accounts posit that childrens early language reflects the presence of syntactic abstractions, constructivist approaches instead emphasize gradual generalization over frequently-heard forms. Here we develop a Bayesian statistical model that measures the degree of abstraction implicit in childrens early use of the determiners "a" and "the." Our work reveals that many previously-used corpora are too small to adjudicate between these theoretical positions. Several datasets, including the Speechome Corpus—a new ultra-dense dataset for one child—show evidence of low initial levels of productivity and higher levels later in development, however. These findings are consistent with the hypothesis that children lack rich grammatical knowledge at the outset of language learning, but rapidly begin to generalize on the basis of structural regularities in their input.

Keywords: language acquisition; Bayesian models; corpus linguistics

One of the most astonishing parts of children's language acquisition is the emergence of the ability to say and understand things that they have never heard before. This ability, known as *productivity*, is a hallmark of human language (von Humboldt, 1970/1836; Hockett, 1959). Indeed, adults' linguistic representations are almost universally described in terms of syntactic abstractions such as "determiner," "verb," and "noun phrase" (e.g., Chomsky, 1981; Sag, Wasow, & Bender, 1999). But do these same adult-like abstractions guide how children produce and comprehend language?

Some researchers have suggested a *generativist* view of syntactic acquisition: adult-like abstractions guide children's comprehension and production from as early as it can be measured (Pinker, 1984; Valian, 1986; Yang, 2013). Others have argued that adult-like syntactic categories—or at least their guiding role in behavior—emerge gradually, with the accumulation of experience. On such *constructivist* views, children's representations progress over time from memorized multi-word expressions to specific item-based constructions and eventually generalize to abstract combinatorial rules (Braine, 1976; Pine & Martindale, 1996; Pine & Lieven, 1997; Tomasello, 2003).

Here we focus on a key case study for this debate: the emergence of the capacity in English to produce a noun phrase (NP) by combining a determiner (Det, such as "the" or "a") with a noun (N). This capacity is exemplified for adult English by the context-free rule

$$\text{NP} \rightarrow \text{Det N}$$

Using this knowledge, when adult native English speakers hear a novel count noun with "a," e.g. "a blicket," they know that combining the novel noun with "the" will also produce a permissible noun phrase, e.g. "the blicket." Adult-like knowledge and use of this part of English syntax requires the category Det, the category N, and the rule specifying how they are combined.

In recent years, this case study—noun phrase productivity, with a focus on the use

of determiners—has played an increasingly prominent role in the generativist–constructivist child language acquisition debate (Valian, 1986; Pine & Martindale, 1996; Pine & Lieven, 1997; Valian, Solt, & Stewart, 2009; Yang, 2013; Pine, Freudenthal, Krajewski, & Gobet, 2013). Whereas nouns often have referents in the child's environment, the semantic contribution of determiners to utterance meaning is more subtle (Fenson et al., 1994; Tardif et al., 2008). Thus one might expect determiners to be learned late (Valian et al., 2009). Yet children produce them relatively early, and their uses are overwhelmingly correct by the standards of adult grammar. Is this because children deploy adult-like syntactic knowledge, or because they memorize and reuse specific noun phrases, creating the illusion of full productivity?

Experimental methods have been of limited utility in resolving this question. Tomasello and Olguin (1993) found evidence for the presence of a productive object word category in children between 20 and 26 months, presenting objects with nonce labels and eliciting reuse in novel syntactic contexts and morphological forms (using a "wug" test; Berko, 1958). But these data do not resolve the extent to which syntactic abstractions guide their everyday speech. Instead, most work on children's syntactic productivity has relied on observational language samples (Valian, 1986; Pine & Martindale, 1996; Pine & Lieven, 1997; Valian et al., 2009; Yang, 2013; Pine et al., 2013).

Making inferences about children's knowledge from observational evidence is difficult for a number of reasons, however. First, individual child language corpora have typically been small—consisting of weekly or monthly recordings of only a couple of hours. Second, nouns (like other words) follow a Zipfian frequency distribution (Zipf, 1935), in which a small number of words are heard often, but most are heard only a handful of times. As a result, evidence regarding the range of syntactic contexts in which a given child uses an individual noun is weak for most nouns (Yang, 2013). These inferential challenges are sufficiently severe that within the past several years, researchers on

opposing sides of the productivity debate have drawn opposite conclusions from similar datasets (Pine et al., 2013; Yang, 2013). Making progress on children's syntactic productivity requires overcoming these challenges.

Here we present a new, model-based framework for drawing inferences about syntactic productivity, differing from previous work in two critical respects. First, previous approaches assessed productivity via a summary statistic, the *overlap score*, computed from a child language sample. This statistic is difficult to interpret because it may be biased by the size and composition of the sample (discussed below). Here, in contrast, we model productivity as one feature of a model of child language whose parameters can be estimated from a sample and whose overall fit to the data can be assessed. Second, we explicitly model item-based memorization and reuse of specific determiner–noun pairs from caregiver speech in the child's environment as an additional contributor to child language production alongside syntactic productivity. Our framework encodes a continuum of hypotheses ranging between fully productive and fully item-based, and allows us to assess how a child at any given point in development balances these two knowledge sources in their production of determiner–noun combinations.

We apply this model to a wide range of longitudinal corpora of child speech, including the Speechome Corpus (Roy, Frank, DeCamp, Miller, & Roy, 2015), a new high-density set of recordings of one child's early input and productions. Our model reveals that many of the conventional corpora analyzed in previous research (Valian, 1986; Pine & Martindale, 1996; Pine & Lieven, 1997; Valian et al., 2009; Pine et al., 2013) are too small to draw high-confidence inferences. An exploratory analysis of the Speechome data, both denser and from earlier in development, provides evidence for low initial levels of productivity followed by a rapid increase starting around 24 months. Several other datasets provide corroboratory evidence. Contra full-productivity accounts, syntactic productivity is very low in the first months of determiner use in these datasets. At the

same time, the current work constrains the timeline of constructivist accounts. We find a rapid early increase in productivity—in Speechome this increase occurs within a few months of the onset of combinatorial speech, prior to the beginning of many of the datasets that have been used previously to address this question. We conclude by discussing the need for denser datasets to provide conclusive evidence on questions about the roots of syntactic abstraction.

Previous Work and Present Goals

Previous investigations have focused on the overlap score, a summary statistic of productivity (Pine & Martindale, 1996; Pine & Lieven, 1997; Pine et al., 2013). Overlap is calculated from the distribution of determiner–noun pairings in a sample, as the proportion of nouns that appear with both "a" and "the" out of the total number of nouns used with either. While initial investigations suggested that young children use comparatively fewer nouns with both determiners (Pine & Martindale, 1996; Pine & Lieven, 1997), overlap scores are highly dependent on sample size due to the Zipfian distribution of noun frequencies (Valian et al., 2009; Yang, 2013). In addition, this statistic is not well-suited for distinguishing increases in direct experience—greater exposure to the relevant words (e.g., hearing both "the dog" and "a dog" independently and subsequently repeating these, even without abstraction)—from true changes in grammatical productivity (Valian et al., 2009; Yang, 2013).

Two recent investigations have used more sophisticated techniques to address issues of sample size. Yang (2013) constructed a null-hypothesis "full productivity" model in which each noun has the same distribution over determiner pairings (no item-specific preferences) and showed that it predicted overlap score well for six children in the CHILDES database. Pine et al. (2013), in contrast, developed a noun-controlled method for comparing adult and child productivity scores in a given sample, and rejected a

full-productivity null hypothesis. Neither of these methods, however, is well suited to tracking developmental changes in productivity, because of their focus on the overlap score. If item-based knowledge plays a role in children's productions, overlap might increase over time even without any changes in productivity, simply because children have heard more determiner+noun pairs.

Here we take a fundamentally different approach from previous work, to address the challenge of decoupling genuinely productive behavior from what might be expected on the basis of experience. We proceed from the observation that there are two sources of information by which a speaker could know that a particular determiner–noun pair belongs to English, and thus potentially produce it: (1) direct experience with that specific determiner–noun pair and (2) a productive inference using knowledge abstracted from experience with different determiner–noun pairs (and perhaps other input as well). Measuring a given speaker's productivity from corpus data requires assessing the extent to which the speaker's language use reflects productivity above and beyond what can be attributed to direct experience.

We define a probabilistic model of determiner+noun production that considers both knowledge sources. In our model, the contribution of productive knowledge can range along a continuum from none (a child capable only of imitating caregiver input, like an idealized version of an "island" learner as described in Tomasello, 1992) to complete (a "total generalizer" equivalent to the null-hypothesis model in Yang, 2013). Specific model parameters correspond to the contributions of these two information sources, and we use Bayesian inference to infer likely values of these parameters for a corpus sample given both the child's determiner–noun productions and caregiver input. By comparing temporally successive samples for a given child, we can use this model to estimate the child's change in syntactic productivity over time. Because our model is fully Bayesian, we are also able to estimate the level of certainty in our estimates, critically allowing us to

avoid overly confident inferences when data are too sparse.

Method

Model

We model the use of each noun token with a specific determiner as the output of a probabilistic generative process. We assume that each noun has its own determiner preference ranging from 0 (a noun used only with "a") to 1 (a noun used only with "the"). We then explicitly model cross-noun variability by assuming some underlying distribution of determiner preferences across all nouns. Lower cross-noun variability indicates that nouns behave in a more class-like fashion, while higher variability indicates little generalization of determiner use across nouns.

Formally, each noun type can be thought of as a coin whose weight corresponds to its determiner preference. Each use of that noun type with a determiner is thus analogous to the flip of that weighted coin, where heads indicate the use of the definite determiner and tails correspond to the indefinite. A sequence of noun uses are thus draws from a binomial distribution with success parameter μ corresponding to the determiner preference. The determiner preference for each noun is drawn from a beta distribution with mean μ_0 (the underlying "average" preference across all nouns) and scale ν, giving us a *hierarchical beta-binomial* model (Gelman, Carlin, Stern, & Rubin, 2004).[1]

Under this model, a child's determiner productions for each noun she uses are guided by a combination of the two information sources mentioned above—(1) direct experience, and (2) productive knowledge—and the strength of each information source's contribution to the child's productions is determined by a weighting parameter. For (1), a parameter η determines how effectively the child learns from noun-specific determiner

[1]Many readers may be more familiar with the more common parameterization of the beta distribution in terms of shape parameters $\alpha = \mu\nu$ and $\beta = (1 - \mu)\nu$.

productions in its linguistic input; for (2), a parameter ν determines how strongly the child applies productive knowledge of determiner use across *all* nouns. These parameters η and ν do not trade off against each other, but rather play complementary roles in accounting for a child's productions: As η increases, the variability across nouns in a child's determiner productions can more closely match the variability in her input, while as ν increases, the child is increasingly able to produce determiner–noun pairs for which she has not received sufficient evidence from caregiver input.

At the heart of the model are the contributions of direct experience and productive knowledge. Both contribute to the rate at which a child uses "the" as opposed to "a" for each noun. This rate, μ_i, is taken to be beta-distributed and corresponds to a beta-binomial Bayesian update of a prior of mean μ_0 and concentration ν with count data corresponding to the caregiver input, weighted by a factor of η. Thus, larger ν indicates stronger influence from the child's productive knowledge, while larger η indicates that the child learns the noun-specific nuances of caregiver input more effectively. For more details see *SOM: Parameters of Beta-Binomial Model*; Our complete hierarchical Bayesian model and variable definitions are presented in Fig. S1.

Since we lack exhaustive recordings of caregiver input, we treat unrecorded caregiver input as a latent variable drawn from the same distribution as aggregated caregiver input and infer it jointly with model parameters (see *SOM: Details of the Imputation*). The theoretically critical target of inference is ν, the strength of the child's productive knowledge of determiner–noun combinatorial potential, which can range from $\nu = 0$ (an extreme "island" learner whose determiner preference for a given noun is guided exclusively by its direct experience with that noun, and whose noun-specific determiner preferences are likely to be skewed toward 0 or 1) to ν approaching infinity (an extreme over-generalizer who has identical determiner preference for all nouns, Fig. 1A).

We use Markov chain Monte Carlo to infer confidence intervals over η, μ_0, and ν

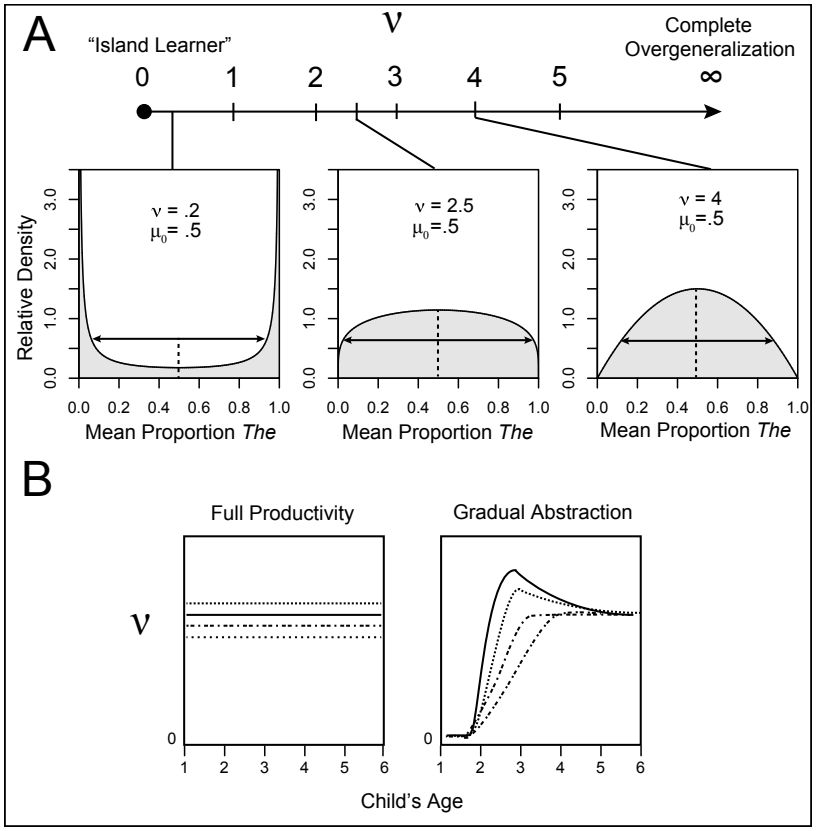


Figure 1. : **A**: Interpretation of the ν parameter, a concise metric of grammatical productivity. At low values of ν, little or no information is shared between nouns. At higher ν values, nouns exhibit more consistent usage as a class, indicating the existence of a productive rule governing the combination of determiners and nouns. μ_0 represents the mean proportion of definite determiner usage across nouns, set here at .5 in all three panels. **B**: Schematized trajectories for the development of grammatical productivity under two competing theories.

from a child's recorded productions and linguistic input. But a single recording of a child typically does not yield high confidence in these estimates because of the relatively low numbers of productions for individual nouns. To overcome this issue, we use two different methods for constructing sufficiently large samples of child and caregiver tokens to evaluate the developmental trajectory of the ν parameter: split-half and sliding window analyses.

First, in the *split-half* analysis, we divide the data for each child into distinct early and late time windows with an equal number of tokens, denoted with the subscripts t_1 and t_2. Separate parameter sets (μ, ν, η) are maintained for the first and second windows; for a given sample from the joint posterior, the changes in parameters from the first window to the second can be calculated as:

$$\Delta\nu = \nu_{t_2} - \nu_{t_1}, \qquad \Delta\mu = \mu_{t_2} - \mu_{t_1}, \qquad \Delta\eta = \eta_{t_2} - \eta_{t_1}. \tag{1}$$

These variables may be treated as targets of inference, over which highest posterior density (HPD) intervals may be computed. Our principal target of inference is $\Delta\nu$, the change in the contribution of productive knowledge to the child's determiner use. This two-window approach maximizes statistical power, but does so at the expense of a detailed time-related trajectory: for those children with longer periods of coverage, this estimate may group together several distinct developmental time periods.

Second, as an exploratory technique, we also use our model to measure finer-grained changes in parameter estimates across development via a *sliding window* approach in which the model is fit to successively later subsections of the corpus of child productions. Each window also includes the corresponding adult productions that occur prior to or during that subsection. In this case we fit the model to successive 1024 token windows of the child's speech, advancing by 256 tokens for each sample. This method yields a higher

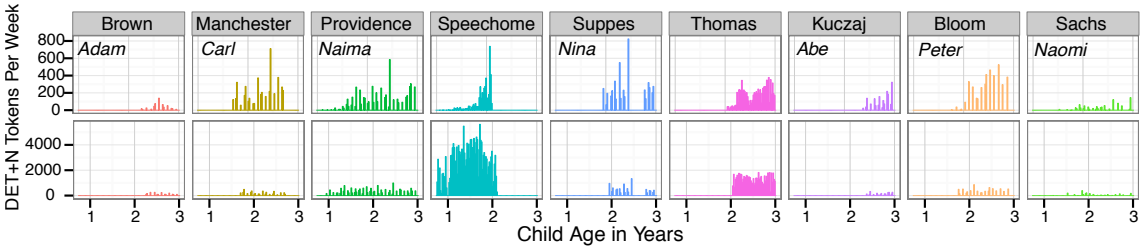


Figure 2. : Number of recorded determiner+noun uses per week for children (top) and corresponding caregivers (bottom) before three years of age for the child with the most determiner+noun pairs for each of the nine corpora analyzed here. Note that the number of weekly observations from children and caregivers are presented on differing scales (0-800 and 0-6,000 respectively).

resolution timecourse than the split-half analysis, though at the expense of less-constrained parameter estimates, especially for the smaller corpora. For more details on inferring model parameters, see *SOM: Model Fitting Procedure*.

Our approach is an example of "Bayesian data analysis" (Gelman, Carlin, Stern, & Rubin, 2003). We create a cognitively interpretable model that captures the spectrum of different hypotheses, from item-based learning to full productivity. We can then infer, for a particular dataset, where on the spectrum the data fall. In a classic predictive model, parameters are fit–or overfit—to some external performance standard. In contrast, our model summarizes a particular aspect of the dataset and gives an estimate of the relative certainty we have in this summary measurement.

Data

We used a large set of publicly-available longitudinal developmental corpora of recordings of children and their caregivers from the CHILDES archive (MacWhinney, 2000). Four of these corpora have been examined previously for early evidence of grammatical productivity: the Providence Corpus (Demuth, Culbertson, & Alter, 2006),

the Manchester Corpus (Theakston, Lieven, Pine, & Rowland, 2001), the Brown Corpus (Brown, 1973), and the Sachs Corpus (Sachs, 1983). We additionally analyze four single-child corpora: Bloom (Bloom, Hood, & Lightbown, 1974), Kuczaj (Kuczaj, 1977), Suppes (Suppes, 1974), and Thomas (Lieven, Salomo, & Tomasello, 2009). These eight corpora yield usable data for a total of 26 children. While high-density data with rich annotations exist for all of these corpora, coverage starts in most cases well after the onset of combinatorial speech and is sparse under two years of age, the time interval necessary for characterizing initial levels of grammatical productivity.

To address these shortcomings, we additionally analyze the densest longitudinal developmental corpus in existence, the Speechome Corpus (Roy et al., 2015). The Speechome Corpus covers the period 9 through 25 months in the life of a single child, and contains video and audio recordings of nearly 70% of the child's waking hours, with transcripts for a substantial portion of these (Vosoughi & Roy, 2012). While transcription of the Speechome Corpus is a work in progress, the version used here contains approximately 4,300 noun phrases with articles produced by the child before 25 months of age and includes dense coverage of child-accessible caregiver speech, with some 196,300 noun phrases in the same time period. The Speechome Corpus supports more detailed inferences about developmental timecourse in the second year of life. The Speechome Corpus is also distinguished in the quantity of child-available adult speech, with nearly 80% more caregiver tokens than the next best-represented child, Thomas. Fig. 2 shows comparative densities for adult and child determiner+noun pairs for the child with the most data in each corpus.[2]

We assess our model using seven different methods for extracting determiner+noun

[2]Datasets that capture large amounts of an individual family's experience like Speechome pose unique privacy risks. In order to share reproducible data while maintaining privacy, we are distributing determiner+noun count data from the Speechome corpus while obfuscating the identities of the specific nouns the child produced.

data from each corpus. These data treatments reflect a range of assumptions regarding the availability of phrase structure for identifying which noun corresponds to each determiner, whether information can be shared between morphologically inflected forms, and whether the child is considering only singular forms in the language. In the absence of reliable morphological tags, the Thomas and Speechome corpora were assessed on four data treatments each. For additional technical details refer to *Supplementary Material: Data Preparation*. We have made available model code, noun-anonymized Speechome data, and auxiliary code necessary to reproduce our research in a public GitHub repository accessible at `https://github.com/smeylan/determiner_learning`.

Results

The two hypotheses represented in the literature—full productivity or gradual abstraction over item-based knowledge (Fig. 1B)—make contrasting predictions regarding initial productivity and the effects of developmental change. Full productivity predicts a nonzero initial level combined with a negligible effect of developmental time—productivity does not increase with exposure to more data. Gradual abstraction over item-based knowledge, in contrast, predicts near-zero initial productivity indicating the absence of syntactic category knowledge in the earliest productions, and a positive relationship with developmental time corresponding to the gradual induction of abstract categories throughout childhood.

Split-Half Method

To test for changes in productivity, we assess the null hypothesis that 0 (no change) is within the 99.9% HPD interval for the posterior estimate of $\Delta\nu$, the difference in ν estimates between the first and second half of tokens each child. (We use the 99.9% criterion because of the large number of independent comparisons implied by this analysis—one for each of the 27 children.) By this standard, only one child (Speechome,

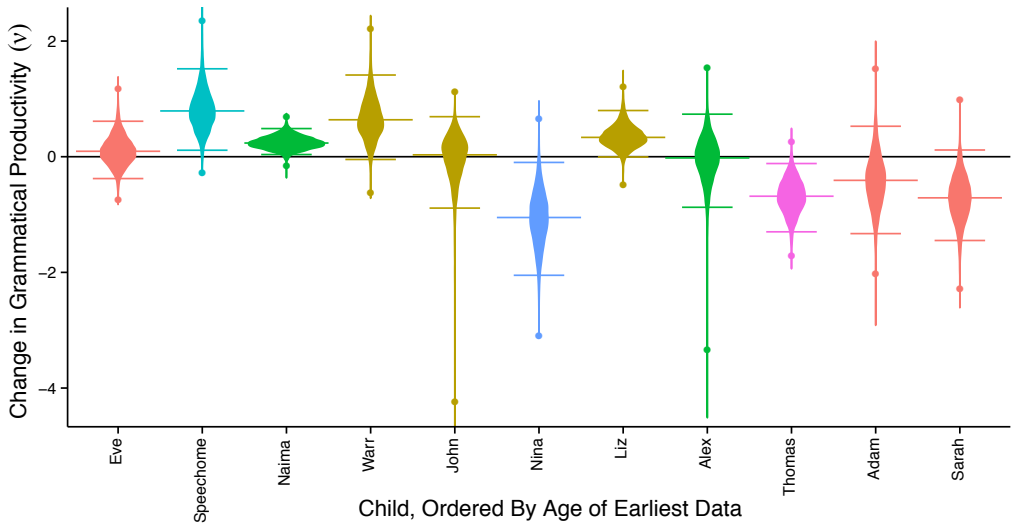


Figure 3. : Posterior estimates for change in productivity between first and second half of children's corpora ($\Delta\nu$) for each child (n = 11). Longest horizontal lines indicate the median of the posterior, and shorter horizontal lines the 95% HPD. Points indicate the 99.9% HPD. The remainder of the children (n=16) are not displayed on the basis of poorly constrained posteriors (99.9% HPD for ν outside [0,3] for either time period).

in 3 of 4 data treatments) shows a significant increase in productivity (Fig. 3 and Fig. 4). The remaining data treatment for Speechome is strictly positive within the 95% HPD interval. Three other children—Liz (in 3 of 7 data treatments), Naima (1 of 7 treatments) and Warr (1 of 7 treatments)—have at least one data treatment where change is strictly positive within the 95% HPD. These findings suggest some early increases in productivity. Results across all seven data preparations are presented in Figure S4.

We also found apparent *decreases* in grammatical productivity for several of the older children. Thomas (2 of 4 treatments within the 99.9% HPD interval, 1 in the 95% HPD interval), Sarah (1 in the 99.9% and 4 in the 95% HPD interval), and Nina (2 in the 99.5% and 2 in the 95 % HPD interval) show strictly negative changes. The timing of these decreases are consistent with a phase of overregularization, during which they are

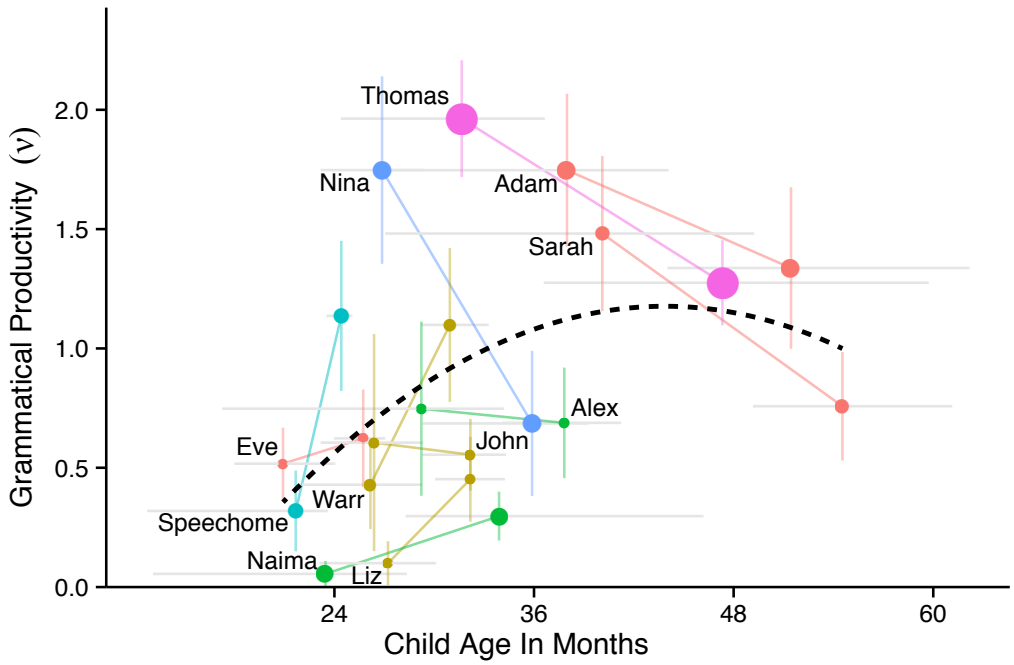


Figure 4. : The inferred developmental trajectory for determiner productivity, ν, across children ($n = 11$). Each line shows a two-point productivity trajectory for a single child, plotted by age in months. Marker size corresponds to the number of child tokens used for each child. Gray horizontal lines indicate the temporal extent of the tokens used to parameterize the model at each point; vertical lines indicate the SD of the posterior. The best fitting quadratic trend is shown as a dashed black line.

more willing to use determiner noun-combinations that are rare or unattested in adult speech like *a sky*, followed by a decrease towards adult-like levels. Consistent with this hypothesis, increases in ν tended to occur in datasets from younger children (p=0.009 by rank sum test on the children in Fig. 3).

Together these results are broadly consistent with constructivist hypotheses, in that we find minimal evidence of productivity in the earliest multiword utterance coupled with a development-related increase in productivity soon thereafter. However, our results

deviate slightly from the proposal of gradual emergence of abstract schema from item-specific exemplars, as set forth in (Abbot-Smith & Tomasello, 2006). The possibility of a decrease in determiner productivity later in development suggests that while children may construct abstract generalizations from their input, they may also use input later in development to constrain overly general abstract schema (along the lines schematized in Figure 1B, right, top two trajectories).

Our model is defined independently from overlap score, the primary measure of productivity used in previous literature. We can take advantage of this independence to use overlap as a model validation method. Although a simple overlap measure is not useful for characterizing productivity and comparing *across* children, we can use it to validate our model *within* individuals. We do this by sampling new simulated determiner productions from the fitted model's distribution on child determiners for each time window, computing overlap, and then comparing the results to the empirical values from that same child. Empirical overlap falls within the 95% range of simulated overlap scores for all children, validating the model's overall fit to the data. For additional details see *Supplementary Material: Results*.

Sliding Window Method

The higher temporal resolution sliding window method reveals changes in grammatical productivity consistent with the split-half analysis, with major increases in productivity for Speechome and Warr and major decreases for Thomas, Adam, and Sarah (Figure 5, column 1). The sliding window models also reveal significant variability in ν not related to age (e.g., Naima from the Providence Corpus). In addition, using the same validation technique described above, simulated overlap from sliding window estimates was strongly correlated with empirical values (Pearson's r of 0.940 – 0.951 across data treatments).

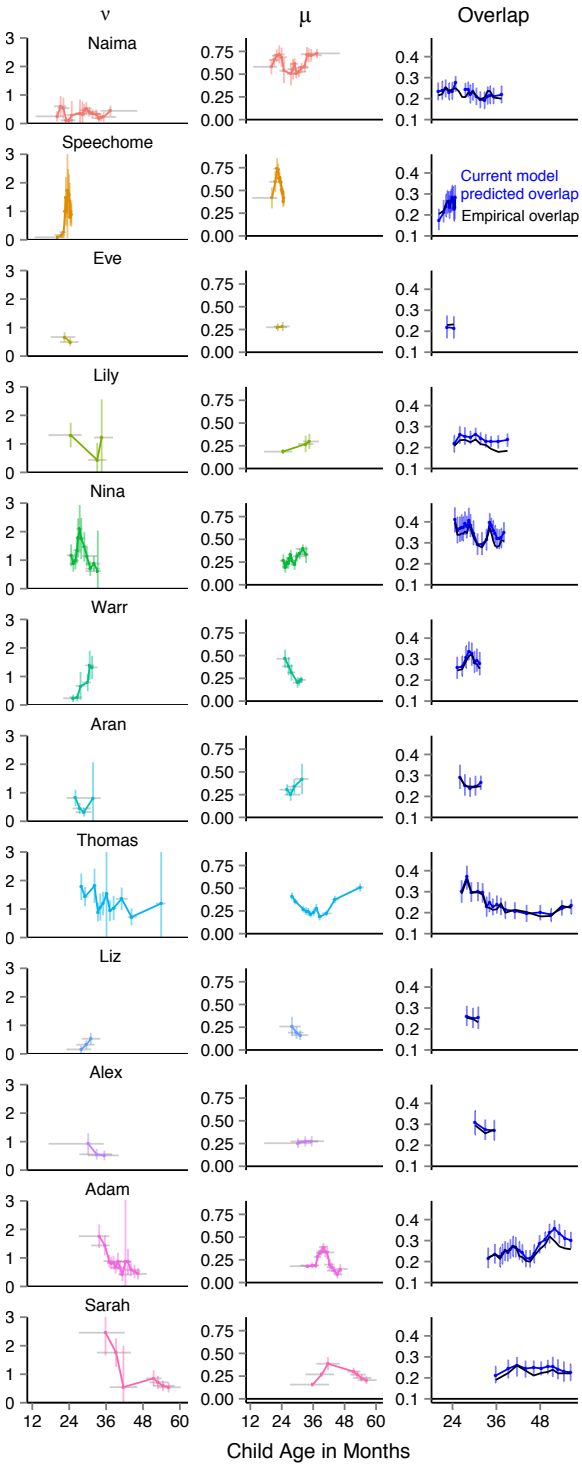


Figure 5. : Child determiner productivity ν, child mean determiner preference μ, and predicted and empirical overlap scores for the 11 children presented in the split half analysis. Vertical lines show the 99% HPD for ν, μ, and overlap predicted by the current model. Horizontal lines indicate the temporal extent of the tokens used to fit the model at each point.

Discussion

The model-based statistical approach presented here for analyzing child language is the first method that allows the respective contributions of productivity and item-based knowledge to be teased apart. Our analysis reveals two key findings. First, children's syntactic productivity changes over development. Several of the youngest children show increases in productivity, with evidence strongest in the largest dataset, Speechome. In addition, some older children show decreases in productivity. This trend might suggest a period of particularly strong generalization followed by a retreat, similar to the pattern observed in morphological domains (e.g., Rumelhart & McClelland, 1985; Pinker, 1991), as well as verb argument structures (Bowerman, 1988; Ambridge, Pine, & Rowland, 2011).

Second, for the majority of children, our model placed wide confidence intervals on productivity estimates, indicating that the available data were likely not sufficient to draw precise developmental conclusions. The data for these children typically included a maximum of one hour per week of transcripts; furthermore, most of the child productions in these datasets were collected after the child's second birthday. If adult-like categories are constructed early—soon after the onset of word combination—many of these datasets begin too late to provide decisive evidence regarding the trajectory of early development. The trend line obtained in Figure 4 is suggestive rather than conclusive; additional datasets would be required to test whether the pattern is robust within the developmental trajectory of a single child. These results underscore the critical importance of dense, naturalistic data for understanding the development of linguistic knowledge in early childhood.

Debates about the emergence of syntactic productivity have typically oscillated between two poles: Immediate, full productivity early in development, or accumulation of item-specific knowledge with gradually increasing levels of productivity. Our approach parameterizes the space of models between these poles. In the future it can be adapted to

characterize productivity in other simple morphosyntactic phenomena and in other languages. In the key case study of English determiner productivity, applying our model to new, dense data yields support for constructivist accounts and further constrains the developmental timeline within these accounts. While children's earliest multiword utterances may be island-like, grammatical productivity emerges rapidly thereafter.

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, *23*, 275–290.

Ambridge, B., Pine, J., & Rowland, C. (2011). Children use verb semantics to retreat from overgeneralization errors: A novel verb grammaticality judgment study. *Cognitive Linguistics*, *22*(2), 303–323.

Berko, J. (1958). The Child's Learning of English Morphology. *Word*, 150–177.

Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: If, when and why. *Cognitive Psychology*, *6*, 380-420.

Bowerman, . (1988). The'no negative evidence'problem: How do children avoid constructing an overly general grammar? In *Explaining language universals* (pp. 73–101). Basil Blackwell.

Braine, M. (1976). Children's first word combinations. *Monographs of the Society for Research in Child Development*, *41*(1).

Brown, R. (1973). *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris Publishers.

Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language & Speech*, *49*(2), 137–173.

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D., Pethick, S., ... Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 1–185.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2003). *Bayesian data analysis*. CRC press.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Hockett, C. (1959). Animal languages and human languages. *Human Biology*(31), 32-39.

Kuczaj, S. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, *16*, 589–600.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, *20*(3), 481-508.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates.

Pine, J., Freudenthal, D., Krajewski, G., & Gobet, F. (2013). Do young children have adult-like syntactic categories? Zipf's Law and the case of the determiner. *Cognition*, *127*, 345-360.

Pine, J., & Lieven, E. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, *18*(2), 123–138.

Pine, J., & Martindale, H. (1996). Syntactic categories in the speech of young children: The Case of the Determiner. *Journal of Child Language*, *23*(2), 369–395. doi: 10.1037/0012-1649.22.4.562

Pinker, S. (1984). *Language Learnability and Language Development*. Cambridge University Press.

Pinker, S. (1991). Rules of language. *Science*, *253*, 530–535.

Roy, B., Frank, M., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663-12668.

Rumelhart, D., & McClelland, J. (1985). On learning the past tenses of English verbs. In *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2). Cambridge, MA: MIT Press.

Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference

in parent–child discourse. In *Children's Language* (Vol. 4). Lawrence Erlbaum Associates.

Sag, I., Wasow, T., & Bender, E. (1999). *Syntactic theory: A formal introduction.* Stanford, CA: CSLI Press.

Suppes, P. (1974). The semantics of children's language. *American Psychologist*, *29*, 103-114.

Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N., & Marchman, V. (2008). Baby's first 10 words. *Developmental Psychology*, *44*, 929.

Theakston, A., Lieven, E., Pine, J., & Rowland, C. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, *28*, 127–152.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Tomasello, M., & Olguin, R. (1993). Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development*, *8*(4), 451–464.

Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology*, *2*, 562–579.

Valian, V., Solt, S., & Stewart, J. (2009). Abstract categories or limited-scope formulae? The case of children's determiners. *Journal of Child Language*, *36*, 743–778.

von Humboldt, W. (1970/1836). *On language: On the diversity of human language construction and its influence on the mental development of the human species.* Cambridge University Press.

Vosoughi, S., & Roy, D. (2012). An automatic child-directed speech detector for the study of child language development. In *Proceedings of Interspeech.*

Yang, C. (2013). Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*.

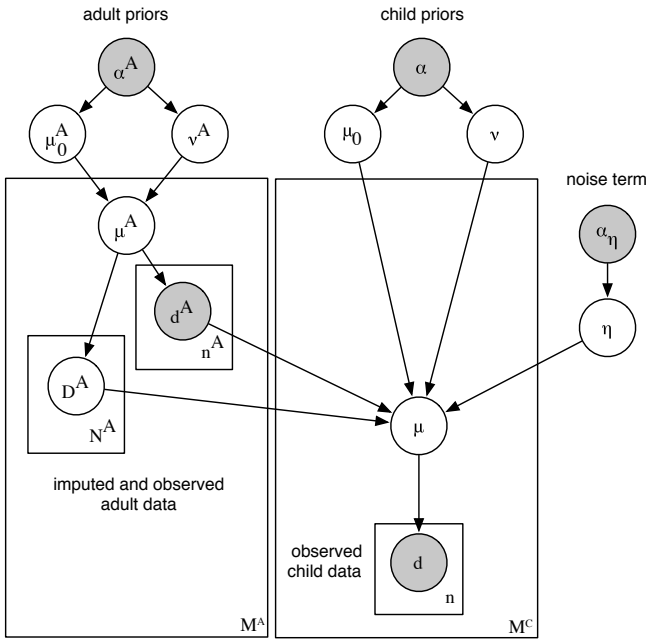
Zipf, G. (1935). *The psycho-biology of language: An introdution to dynamic philology.* MIT Press.

Supplementary Material

Model

Parameters of the Beta-Binomial Model. The rate at which a child uses "the" rather than "a" for each noun i, is treated as a beta-distributed random variable, μ_i. μ_i has mean $\frac{\mu_0\nu+\eta\left(r_i^A+R_i^A\right)}{\nu+\eta\left(n_i^A+N_i^A\right)}$ and concentration $\nu+\eta\left(n_i^A+N_i^A\right)$, where the child has experienced r_i^A researcher-observed and R_i^A researcher-unobserved uses of noun i with "the" and, respectively, $n_i^A-r_i^A$ and $N_i^A-R_i^A$ with "a." μ_0 and ν describe the prior over determiner preferences across all nouns. Specifically, μ_0 indicates the mean determiner preference and ν indicates the concentration (higher values imply that μ_i values are closer to μ_0). η mediates how effectively the child learns from caregiver input the noun-specific determiner preference for each noun. See Figure S1 for the complete graphical model.

Details of the Imputation. In our model the child learns from the totality of the linguistic input in his or her lifetime, of which the caregiver speech in our datasets represents only a sample. A side effect of Bayesian inference in our model is the imputation of unobserved caregiver input—D^A in Fig. 1. For a window starting at time t and ending at time t', we estimate the child's total lifetime number of {*a*,*the*}+noun input tokens from birth through t' based on a rate of 15 million total words of input per year (Hart & Risley, 1995; Mehl, Vazire, Ramírez-Esparza, Slatcher, & Pennebaker, 2007; Roy, Frank, & Roy, 2009) and 20 determiner–noun pairs per 1,000 words (Godfrey, Holliman, & McDaniel, 1992), and assume that nouns occur in the same relative frequencies in the observed and unobserved portions of this total lifetime input. As can be seen in the graphical model in Fig. 1, inferences about the distribution of determiners for noun i in unobserved caregiver input is constrained by three information sources: Observed caregiver utterances involving noun i, observed caregiver utterances involving *other* nouns, which carry information about the "top-level" caregiver determiner preference (modeled



Model equations for noun i:

$$\mu_i^A \sim \text{Beta}\left(\mu_0^A, \nu^A\right)$$
$$\mu_i \sim \text{Beta}\left(\frac{\mu_0\nu + \eta\left(r_i^A + R_i^A\right)}{\nu + \eta\left(n_i^A + N_i^A\right)}, \nu + \eta\left(n_i^A + N_i^A\right)\right)$$
$$r_i^A \sim \text{Binom}\left(n_i^A, \mu_i^A\right)$$
$$r_i \sim \text{Binom}\left(n_i, \mu_i\right)$$
$$R_i^A \sim \text{Binom}\left(N_i^A, \mu_i^A\right)$$

Variable definitions:

ν	Strength of child's generalized knowledge regarding determiner preference
μ_0	Child's generalized determiner preference
μ	Child's noun-specific determiner preferences
η	Noise parameter indicating child's effectiveness at learning noun-specific determiner preferences from input
ν^A	Dispersion of caregivers' noun-specific determiner preferences
μ_0^A	Caregivers' generalized determiner preference
μ^A	Caregivers' noun-specific determiner preferences
α	Uninformative prior over μ_0, ν
α_η	Uninformative prior over η
α^A	Uninformative prior over μ_0^A, ν^A
d	Child-produced determiner-noun pairs observed in dataset (comprised of r_i "the" instances and $n_i - r_i$ "a" instances for noun i)
d^A	Caregiver-produced determiner-noun pairs observed in dataset (comprised of r_i^A "the" instances and $n_i^A - r_i^A$ "a" instances for noun i)
D^A	Caregiver-produced determiner-noun pairs *not* observed in dataset (comprised of R_i^A "the" instances and $N_i^A - R_i^A$ "a" instances for noun i)

Figure 1. : Graphical representation of our model. Variables with A superscripts (e.g., μ^A) are "adult" (caregiver) parameters; unsuperscripted variables are child parameters. Shaded nodes indicate observed data (adult and child determiner+noun productions d^A and d) or uninformative priors set by the researcher (α^A, α, and α_η).The M^A and M^C plates correspond to noun types used by the caregiver(s) and the child, respectively; the N^A plate corresponds to adult imputed uses of a given noun, n^A to observed adult uses, and n to observed child uses.

as a beta prior with mean μ_0^A and concentration ν^A on noun-specific caregiver determiner preferences), and observed child utterances, which are in part guided by caregiver input.

Model Fitting Procedure. We implemented this model using JAGS for Markov chain Monte Carlo based Bayesian inference (Plummer, 2003). For each model, we took 5 chains of 5000 samples after a burn-in of 2000 adaptive samples and 2000 updates, with thinning of 5 samples (yielding 1000 samples per chain, and 5000 samples total). If the Gelman and Rubin Diagnostic—that the 99th percentile of the potential scale reduction factor, $\hat{R}$ was below 1.1, we considered the model to have converged (Gelman, Carlin, Stern, & Rubin, 2004), otherwise we ran the chains until convergence in 1000 sample increments. If the model did not meet these convergence criteria by 20,000 samples (100,000 without thinning), we report it as non-converging. Low autocorrelation and good mixing were confirmed through spot visual inspection.

To determine the expectation and distribution of overlap scores predicted by our fitted model for a given child's productions in some time window where each noun i is observed N_i times with either *a* or *the*, we first draw a sample vector of noun-specific child determiner preferences $\{\hat{\mu}\}$ from our MCMC-chain approximation to the posterior over $\{\mu\}$, and then draw for each noun i a new binomially distributed sample of size N_i with mean $\hat{\mu}_i$. The proportion of such samples with at least one instance of both *a* and *the* constitutes a single predicted overlap score for that window. By repeating this process over many sample vectors from the chain, we approximate the posterior predictive distribution on the overlap score for that window, and use it to compute expectations and corresponding HPD intervals.

Data Extraction and Preparation

Corpus work at the scale we describe here is necessarily noisy: poor audio quality, annotator idiosyncrasies, and probabilistic methods for extracting hundreds of thousands

of tokens mean that the input to our model inevitably deviates from an ideal data source. Our strategy was thus to test our model across a variety of data preparations to confirm that deviations are of acceptably small magnitude to provide reliable input to the model; indeed, none of the analyses provide us with evidence of systemic problems that might compromise the integrity of our results.

Data Sources. Transcripts for eight developmental corpora (Brown, 1973; Suppes, 1974; Bloom, Hood, & Lightbown, 1974; Kuczaj, 1977; Sachs, 1983; Theakston, Lieven, Pine, & Rowland, 2001; Demuth, Culbertson, & Alter, 2006; Lieven, Salomo, & Tomasello, 2009) were downloaded from the CHILDES project at `childes.psy.cmu.edu`. Utterances from these children ($n = 26$) and their respective caregivers—typically mothers, but also including fathers—were extracted from CHAT-formatted transcripts (MacWhinney, 2000). These specific corpora were selected because they provide longitudinal coverage within the developmental time period of interest, contain annotated samples of both child and caregiver speech, and in many cases have been used extensively in previous research on grammatical productivity.

To test the model on higher-density data than the corpora available in the CHILDES database, we additionally extracted noun phrases from a ninth corpus, the Speechome Corpus (Roy, Frank, DeCamp, Miller, & Roy, 2015). This annotated corpus spans the 9 through 24 month age range of a child's life (n=1). Embedded cameras and microphones located throughout the child's house were used to achieve an unprecedented level of coverage of language learning in a naturalistic context (Figure S2). Annotating the Speechome Corpus was accomplished using new, semi-automatic tools designed for speed and efficiency. Speech from approximately 10 hours per day of raw audio was transcribed using BlitzScribe, an automated system that uses machine learning techniques to segment speech, assign speaker identities, and provide a first pass on transcription. An estimated 72% (3,618 of 5,000 utterances) of caregiver speech from a balanced sample

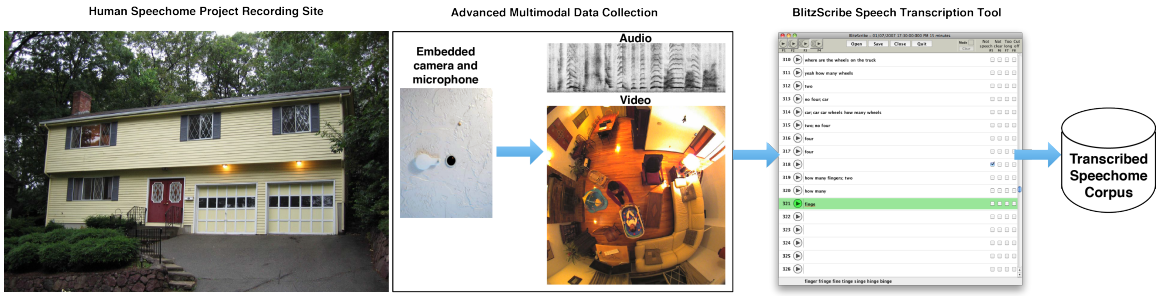


Figure 2. : The Human Speechome Project consists of dense, longitudinal data collection from cameras and microphones embedded in each room of a single child's house. These recordings were transcribed using the custom BlitzScribe transcription tool. The corpus consists of more than 8 million words of transcribed speech and 200,000 hours of audio and video, comprising more than 200 terabytes of media.

across time is child-directed, while the remainder is spoken in the presence of the child but not to the child (Vosoughi & Roy, 2012).

Data Preparation. Determine-noun pairs were extracted from the corpora using three alternative processing pipelines. In the first pipeline ("CLAN"), we extracted determiner and noun pairs from all corpora with CHILDES-compliant annotations using either manually-annotated or, more commonly, machine-generated dependency parses (Sagae, Davis, Lavie, MacWhinney, & Wintner, 2010). While CLAN is a simple rule-based dependency parser, it incorporates significant domain knowledge and uses special annotations available in CHILDES-formatted files in generating parses. As such, it avoids some of the pitfalls that undermine statistical part-of-speech taggers, often trained on adult speech, when run on samples of early child language.

The two largest datasets, Thomas and Speechome, lack canonical CHILDES annotation, and can only be processed using a statistical part of speech tagger. For this reason we employed two alternative pipelines for extracting determiner+noun pairs, both using a state-of-the-art statistical part-of-speech tagger (Toutanova, Klein, Manning, &

Singer, 2003). Determiners that appear without nouns because of interruptions in conversational turn-taking or speech errors were discarded. When the POS tagger identified a series of nouns, we took the first noun as the head of the phrase (the "FN" pipeline) or the last noun as the head of the phrase (the "LN" pipeline).

For all three data extraction pipelines, unrecognizable nouns ("xxx" and "yyy" in CHILDES-formatted files), proper names,[1] and types shorter than three characters were discarded. Both extraction methods accommodate words intervening between the determiner and noun (e.g. an adjective).

The correct treatment of grammatical variants of similar nouns is not immediately obvious. For example, should a model of determiner productivity track separate counts for "dog" and "dogs," or should these be merged into counts for a single noun? For the CLAN extraction pipeline, we produced three variants of the determiner+noun pairs for each CHILDES dataset. The "Complete" morphology treatment maintained separate counts for all variants; for example, "dog," "doggy," and "dogs" were treated as separate nouns, and their counts were tracked separately. In the "Lemmatized" morphology treatment, records were merged by the lemmatized stem—tokens for any of the three above noun types would be counted as the noun "dog". In the "Singulars" treatment, only singular, unmarked nouns were kept (i.e. counts for "dogs" and "doggy" were discarded). Because the Lemmatized morphology treatment requires morphological parses of the nouns from the CLAN-parsed files, only the Complete and Singulars morphology treatments were available for the LN and FN pipelines.

The combination extraction pipelines and morphology treatments produced seven datasets for each child with fully compliant CHILDES-annotated data, and four datasets for the remaining datasets (Speechome and Thomas). These include 1: Complete-FN, 2:

[1]While proper names are generally unlikely to prepended by a determiner, there are many exceptions, including family names ("The Johnsons"), toponyms ("The Gambia","The Hamptons"), historical eras ('The Great Depression'), and publications ("The New York Times").

Complete-LN, 3: Complete-CLAN, 4: Lemmatized-CLAN, 5: Singulars- FN, 6: Singulars-LN (the data preparation presented in the main text), and 7: Singulars- CLAN . We conduct our model-based analysis on all available variants for each child, but stress in the main text the results of the model run on singular nouns from the LN extraction pipeline for both consistency with previous work (Yang, 2013) and high accuracy and precision when compared with gold-standard manual annotation (described below). Descriptive properties for all datasets (LN-Singulars treatment) are provided in Table 1.

Extraction Procedure Validation. To test the accuracy of the automated extraction pipelines, we compared the lists of identified determiner+noun tokens (before filtering by morphological criteria) with a gold-standard set identified by human annotators. Three paid annotators on Amazon Mechanical Turk found determiner+noun pairs in the first 1000 lines in the first and last corpora for three children: Alex from the Providence corpus, Eve from the Brown Corpus, and Warr from the Manchester Corpus. Discrepancies between annotators were resolved by majority rule.

The three automated extraction pipelines generally provide similar lists of determiner+noun pairs compared to the manual annotations (Table 2). Both the CLAN and LN extraction pipelines outperform the FN extraction stack in terms of recall on the twelve transcripts ($p = .012$ and $p = .011$ respectively, per one-tailed Wilcoxon signed rank tests[2]). The LN extraction pipeline outperforms the FN extraction stack on precision as well ($p = .005$ following the same test).

Imputing Determiner+Noun Counts In Adult Speech. For the imputation procedure in the model, unigram probabilities for nouns in the CHILDES American English datasets were obtained by counting all nouns used with a definite or indefinite determiner by maternal or paternal caregivers in all CHILDES American English corpora as of December

[2]Normal approximations of p values were computed using a continuity correction; zero differences were discarded before ranking absolute differences.

Corpus	Child	Age Range Yr;Mo	Distinct Days	Interval in Days	Child Tokens (Types)	Caregiver Tokens (Types)	Child % After Filter
Bloom	Peter	1;9–3;1	20	492	4,357 (540)	7,824 (731)	82.2
Brown	Adam*	2;3–5;2	53	1,070	6,370 (911)	5,852 (1,005)	78.4
Brown	Eve*	1;6–2;3	10	275	1,304 (332)	2,890 (483)	70.6
Brown	Sarah*	2;3–5;1	131	1,037	3,958 (773)	8,454 (1,181)	82.1
Kuczaj	Abe	2;4–5;0	190	972	6,360 (1,070)	4,935 (1,070)	81.0
Manch.	Anne	1;10–2;9	31	336	1,515 (369)	6,514 (711)	79.5
Manch.	Aran	1;11–2;10	33	340	2,194 (419)	8,168 (996)	77.3
Manch.	Becky	2;0–2;11	33	338	1,787 (424)	4,335 (638)	75.8
Manch.	Carl	1;8–2;8	33	364	4,392 (410)	4,206 (516)	71.0
Manch.	Domin	1;10–2;10	35	363	467 (147)	4,752 (532)	81.9
Manch.	Gail	1;11–2;11	34	362	1,145 (386)	4,404 (870)	79.1
Manch.	Joel	1;11–2;10	35	339	1,429 (402)	4,694 (846)	78.1
Manch.	John*	1;11–2;10	32	338	2,081 (363)	4,561 (753)	71.1
Manch.	Liz*	1;11–2;10	34	338	1,632 (348)	3,716 (624)	70.8
Manch.	Nic	2;0–3;0	33	362	936 (279)	5,312 (850)	71.9
Manch.	Ruth	1;11–2;11	33	367	928 (226)	5,377 (696)	81.5
Manch.	Warr*	1;10–2;9	33	340	2,901 (438)	6,748 (833)	73.0
Prov.	Alex*	1;4–3;5	51	759	1,706 (367)	6,618 (1,063)	77.7
Prov.	Ethan	0;11–2;11	50	731	1,750 (570)	10,299 (1,225)	79.3
Prov.	Lily	1;1–4;0	80	1,067	3,425 (864)	19,077 (2,287)	80.7
Prov.	Naima*	0;11–3;10	85	1,062	5,710 (1,030)	18,478 (1,880)	76.9
Prov.	Violet	1;2–3;11	51	1,014	1,325 (428)	6,562 (1,315)	76.0
Prov.	William	1;4–3;4	44	733	1,332 (355)	6,164 (952)	76.1
Sachs	Naomi	1;2–4;9	65	1,304	1,472 (438)	2,784 (634)	71.9
Speech.	Speech.*	0;9–2;1	419	488	4,281 (448)	196,331 (6,212)	71.0
Suppes	Nina*	1;11–3;3	48	489	6,367 (704)	11,830 (878)	70.1
Thomas	Thomas*	2;0–5;0	376	1,076	18,989 (1,870)	110,720 (3,958)	85.5

Table 1: Age range, type and token counts and other properties of corpora analyzed. Counts reflect a data preparation in which only singular nouns are retained and the last noun of any automatically-identified sequence of nouns is assumed to be the head ("Singulars-LN"). Starred children meet the model's convergence criterion in the main analysis (n=11). Child % After Filter indicates the proportion of tokens retained after the application of repetition and imitation filters similar to those used in Yang (2013).

			CLAN		LN		FN	
Transcript	Child	Speaker	Precision	Recall	Precision	Recall	Precision	Recall
First	Alex	Child	—	—	—	—	—	—
		Caregiver	0.93	0.95	0.95	0.90	0.93	0.88
	Eve	Child	0.80	1.00	1.00	1.00	1.00	1.00
		Caregiver	1.00	0.97	0.91	0.91	0.91	0.91
	Warr	Child	1.00	1.00	1.00	1.00	0.92	0.92
		Caregiver	1.00	1.00	0.96	0.96	0.93	0.93
Last	Alex	Child	0.95	0.95	0.76	0.94	0.67	0.70
		Caregiver	0.94	0.93	0.84	0.88	0.75	0.79
	Eve	Child	0.94	0.94	0.93	0.93	0.93	0.87
		Caregiver	0.85	0.92	0.92	0.96	0.92	0.88
	Warr	Child	0.59	0.78	0.92	0.94	0.81	0.83
		Caregiver	0.80	0.89	0.94	0.94	0.84	0.84

Table 2: Performance of the three automated extraction pipelines compared to gold-standard human annotations for six corpus samples. Recall, the proportion of determiner+noun pairs found by the extraction scripts out of those found by human annotators, reflects the completeness of the extraction method. Precision, the proportion of determiner+noun pairs that were found by human annotators out of those found by the extraction script, reflects the number of false positives. Alex (the child) had no determiner+noun pairs in his first transcript.

2013. Imputation data for the Manchester datasets from the CLAN pipeline are from Manchester alone; for the LN and FN pipeline both British English datasets (Manchester and Thomas) were used. Dialectal differences and conflicting orthographic conventions motivated this decision to maintain separate counts for the imputation. Counts used in the Speechome dataset are from that dataset alone. The imputed caregiver count for each noun is defined as $\lfloor p(n)rd \rfloor$, where $p(n)$ the probability of that noun in the relevant dataset (normalized by the total number of nouns), r is the daily rate of caregiver determiner+noun tokens (here 822), and d is the child age in days.

The coverage provided by the Speechome Corpus allows for an evaluation of the estimated daily rate of determiner+noun pairs used in the imputation step. Given a rate

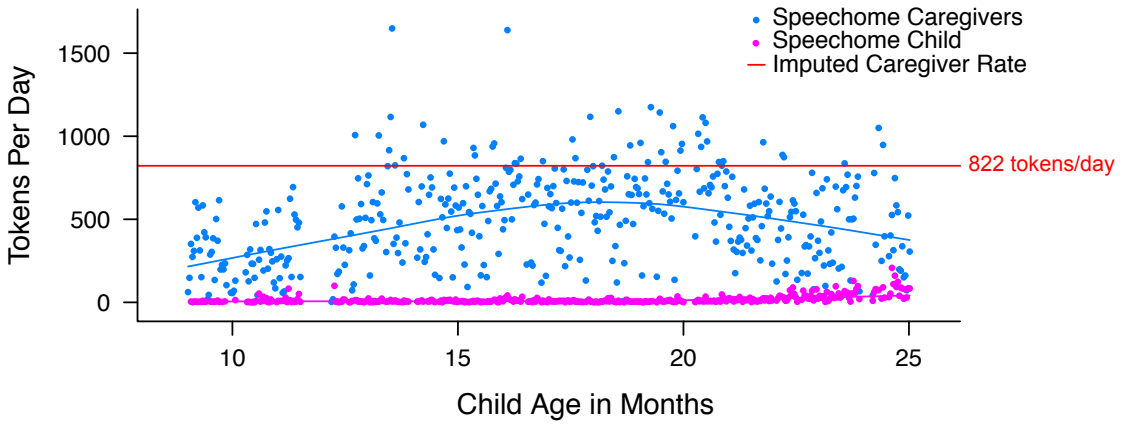


Figure 3. : The observed daily rate of caregiver determiner+noun tokens from the Speechome corpus (blue) is slightly lower than the rate of 822 tokens per day used in the imputation of adult data (marked in red). Loess lines for child and caregiver speech have a span of .67.

of 15 million total words of input per year (Hart & Risley, 1995; Mehl et al., 2007) and 20 determiner+noun pairs per 1,000 words in the Switchboard Corpus (Godfrey et al., 1992), we estimated that a child hears 822 determiner+noun pairs per day. Daily totals of caregiver tokens from Speechome are higher than this estimate (Figure S3). Given that the Speechome corpus is thought to contain approximately 50% of the daily experiences of the target child (∼70% captured, of which ∼70% of the determiner+noun tokens have been annotated), an average of 480 recorded tokens per day corresponds to approximately 960 total determiner+noun tokens per day. We retain the 822 tokens per day as a more conservative estimate.

Additional Analyses of the Speechome Corpus. One potential issue in this particular source data is a bias in the assignment of determiner labels on the annotation process. Portions of the audio from the Speechome dataset were periodically assigned and

transcribed by multiple annotators, providing a way to assess the quality of the annotations in this dataset. Each speech segment has a primary transcript, but may also have a list of alternate transcripts. These alternates can be used to assess quality by computing inter-annotator agreement, the degree to which multiple annotators independently produce the same transcript for a speech segment.

Our analyses are based on a probabilistic model of determiner choice and are thus robust to some level of annotation error. However, we wished to determine if there are biases in annotation errors in that a strong bias in determiner classification toward one of the two determiners could artificially inflate ν. Determiner classification can be technically challenging for automated methods and human annotators alike because it involves distinguishing between highly similar, phonetically reduced segments in fluent speech. Both human annotators and automated methods can take advantage of high-level cues to infer a determiner identity different than that present in the audio signal. Since our main concern here is the child's use of determiners, we selected the subset of child speech segments used in our analyses where alternates were available.

Each alternate transcript is first coded as having either `none`, "a", "the", or `both` determiners present. The latter "`both`" category is required, since in some cases a transcript contains both determiners and it is not always possible to align the determiner to the same target noun used in the primary transcript. The primary transcript, on the other hand, is labeled with the determiner that was linked to the target noun in our analysis (but note that a primary transcript containing multiple determiner+noun pairs may enter into this accuracy calculation multiple times with both "a" and "the" labels.) For a speech segment with $k > 0$ alternates, the counts across the above four categories are accumulated, including the primary transcript category, and normalized by the total number of transcripts $k + 1$. These count vectors are grouped by the primary determiner label, accumulated, and again normalized to yield a confusion matrix shown in Table 3.

Primary Label	None	"the"	"a"	Both	Total Segments
"the"	.22	.74	.03	.00	353
"a"	.28	.03	.67	.03	137

Table 3: Determiner annotation agreement scores for speech segments with multiple transcripts in the Speechome dataset.

The vast majority of alternates agree with the determiner label on the primary transcription; the discrepancies are largely cases where the determiner is dropped. Crucially, there are very few confusions between "the" and "a", and there is no evidence of bias either to switch "the" labels to "a" labels or to switch labels in the opposite direction. Misclassifications remain symmetric over time while the monthly rate of misclassifications decreases with age. This lends support to the analyses and conclusions based on this data.

A second potential issue—in this case also specific to the Speechome dataset given its reliance on automated speaker identification—is the erroneous assignment of caregiver determiner+noun tokens to the child and vice versa. Low precision in automated speaker identification, corresponding to the attribution of caregiver determiner+noun tokens to the child, would inflate the child's ν estimate. To address this concern, two of the co-authors (MCF and BCR) assessed the accuracy of the speaker identification of all child determiner+noun tokens from the Speechome dataset using clips of the original audio data. Of 9,898 machine-identified determiners attributed to the child, 6,875 were confirmed by manual review (Cohen's $\kappa = 0.979$; discrepancies were resolved by discussion). Of these 6,875 tokens, 2,594 were excluded in the LN treatment, and 2,664 in the FN treatment because of dangling determiners, fragmented words, or out-of-vocabulary words. For Speechome, utterances from all adult speakers were aggregated into a single "caregiver" speaker (14% from father, 18% from the mother, 14% from the nanny, 39% attributed to multiple adult speakers, and 13% unsure).

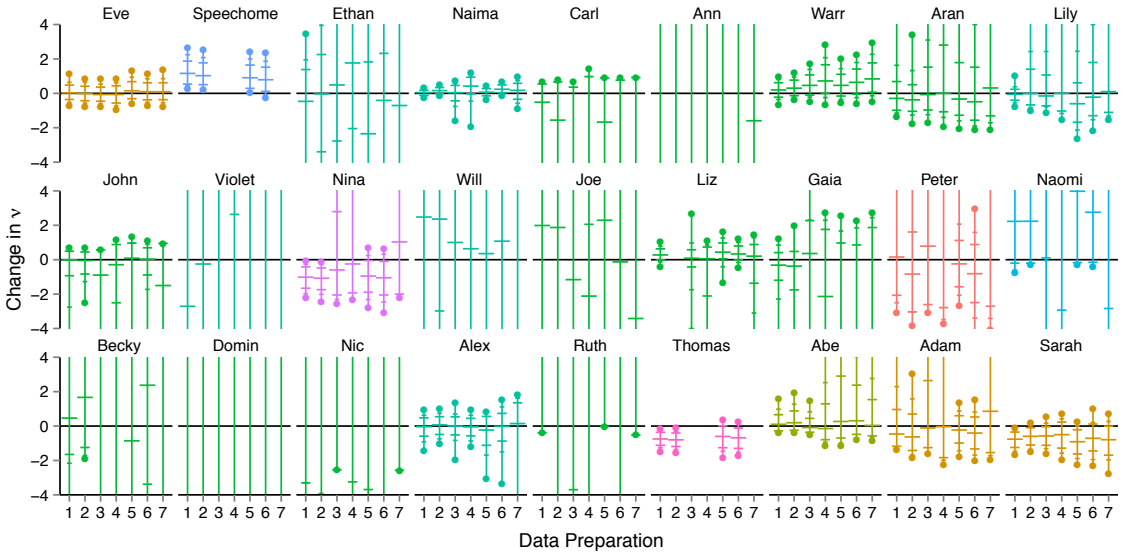


Figure 4. : Posterior distribution of $\Delta\nu$ for all children under all data preparations, 1: Complete-FN, 2: Complete-LN, 3: Complete-CLAN, 4: Lemmatized-CLAN, 5: Singulars-FN, 6: Singulars-LN (the data preparation presented in the main text), 7: Singulars-CLAN . Children are ordered by the mean age of their first interval under the singulars-LN treatment; color indicates the corpus. Vertical lines, from longest to shortest indicate the median of the posterior, the 95% HPD, and the 99% HPD. Points indicate the 99.9% HPD.

Results

Model Convergence. All split-half models converged. Most sliding window models converged (minimum of 267 out of 279 models, in the Complete-LN data preparation).

Predicted vs. Empirical Overlap. Overlap predicted by forward sampling from our model is presented in Table S4. For each data preparation, we performed a Wilcoxon rank sum test comparing the empirical overlap with the overlap computed over forward-sampled det+noun tokens. In no condition did the rank sum test reach significance.

	Current Model			
Data Preparation	r	$RMSE$	< 30 mo.	> 30 mo.
(1) Complete-FN	0.947	0.024	0.023	0.025
(2) Complete-LN	0.941	0.025	0.027	0.024
(3) Complete-CLAN	0.957	0.027	0.027	0.027
(4) Lemmatized-CLAN	0.958	0.032	0.032	0.032
(5) Singulars-FN	0.960	0.028	0.028	0.027
(6) Singulars-LN	0.954	0.029	0.032	0.026
(7) Singulars-CLAN	0.961	0.034	0.036	0.032

Table 4: Pearson's r and root mean squared error for the current model on the split-half data.

Imitation and Repetition Filters for Data. Yang (2013) excluded from analysis child determiner+noun tokens if they were tagged as imitations of the parental speech, as well as within-utterance repetitions by the child. For example, the second instance of "a puzzle" would be discarded if the child said "a puzzle, a puzzle;" if the parent had said "a puzzle" in the preceding utterances *both* would be discarded. A high proportion of repetition and imitation of parental speech on the part of the child could mask initial productivity. On the other hand, such behavior can also be interpreted as genuinely *reflecting* the child's knowledge at that point, in which case excluding such instances from the analysis constitutes an artificial thinning of the data. Constructivist positions assert that the prevalence of rote repetition is itself an important characteristic of children's early speech, rather than noise that must be filtered out to discover some underlying knowledge state (Lieven et al., 2009; Pine & Lieven, 1997). Additionally, imitative and non-imitative uses are hard to distinguish in the real world. Conventions of joint reference in English often lead to cases where two adults use the definite determiner with a noun to

refer to some salient discourse referent; to say that one adult speaker imitates the other in such cases is notably problematic.

We chose not to apply this same filter in our primary analysis in that we consider it to be overly conservative for the reasons outlined above, but we report here the results following an approximation the data preparation in Yang (2013). Because some CHILDES datasets are not annotated with imitation tags and others may follow different classification convention for imitative vs. non-imitative speech, we applied a uniform filter based on repetition of identical tokens in successive utterances. For CHILDES datasets, a child determiner+noun token was omitted from the analysis if it was used by a caregiver in one of the three immediately preceding caregiver utterances in the same file. CHILDES datasets generally lack timestamps, so this method may erroneously exclude child tokens that follow long intervals without annotated material. The Speechome dataset includes high-resolution temporal information that allows for the application of a more fine-grained filter, in which a token was omitted if it occurred within 15 seconds of a caregiver use or another instance of a child use. The proportion of tokens omitted through these filters ranges from 15-30%, with a strong inverse relationship between mean age and the proportion of tokens omitted (see the rightmost column in Table 1).

Crucially, the results of our analysis with these filters are consistent with those presented in the main analysis, though confidence intervals for the estimates are substantially wider (compare Figures 3 and S5). For the Singulars+LN data preparation, only Naima from the Providence corpus reaches the convergence criteria used in the main analysis of 99.9% HPDs for ν in the interval [0,3] in both the first and second half of tokens. Only five children reach convergence when the criteria are weakened to include children with 99.9% HPDs for ν in the interval [0,9]

These children (Speechome, Eve and Adam from the Brown Corpus, Naima from the Providence Corpus, and Thomas) exhibit similar changes in ν from the first to the

second period as in the primary analysis, revealing an overall similar pattern of change (Figure S6). In that HPD intervals are significantly wider, in no case can we reject the null hypothesis of no change between developmental time periods (the decrease for Thomas is marginally significant, however, with no change outside of the 90% HPD). For all children, ν estimates are higher than those reported in the main analysis, suggesting that including repetitions and imitations does indeed produce lower productivity estimates; however, the time-related trends remain robust. We obtain similarly high correlations between predicted and observed overlap (.961–.976 across data preparations), suggesting that this model is equally appropriate for imitation- and repetition- filtered datasets as for unfiltered datasets.

Public distribution of model and data

We have made available model code, noun-anonymized Speechome data, and auxiliary code necessary to reproduce our research in a public GitHub repository accessible at `https://github.com/smeylan/determiner_learning`.

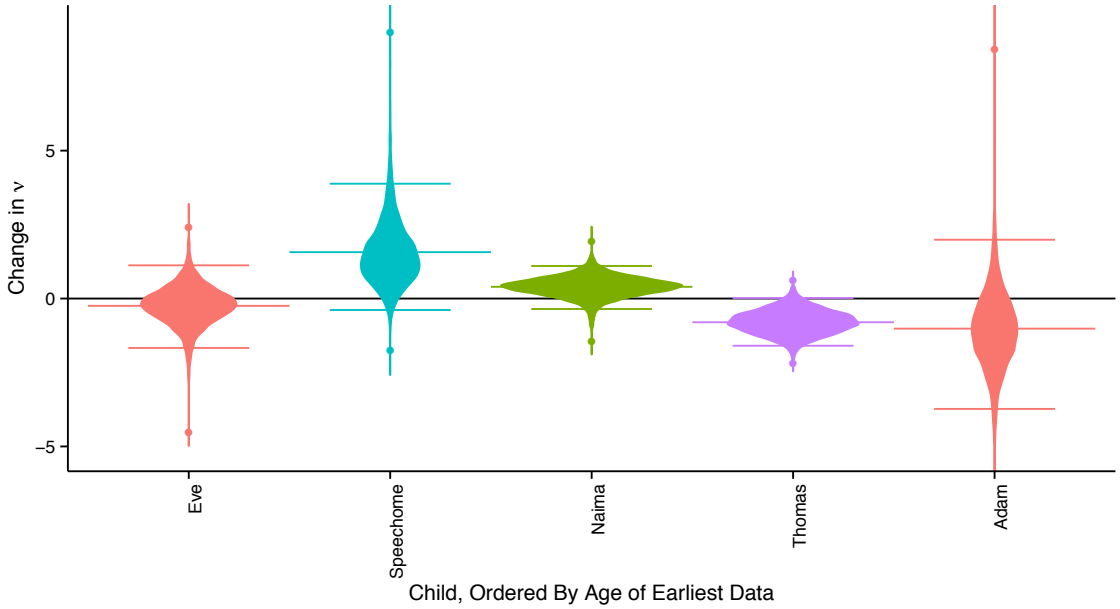


Figure 5. : Posterior estimates for change in productivity between first and second half of childrens corpora ($\Delta\nu$) after applying repetition and imitation filters similar to those used in Yang (2013). Longest horizontal lines indicate the median of the posterior, and shorter horizontal lines the 95% HPD. Points indicate the 99.9% HPD. The remainder of the children (n=22) are omitted on the basis of poorly constrained posteriors relating to small sample sizes (99.9% HPD for ν exceeding [0,9] for either time period).

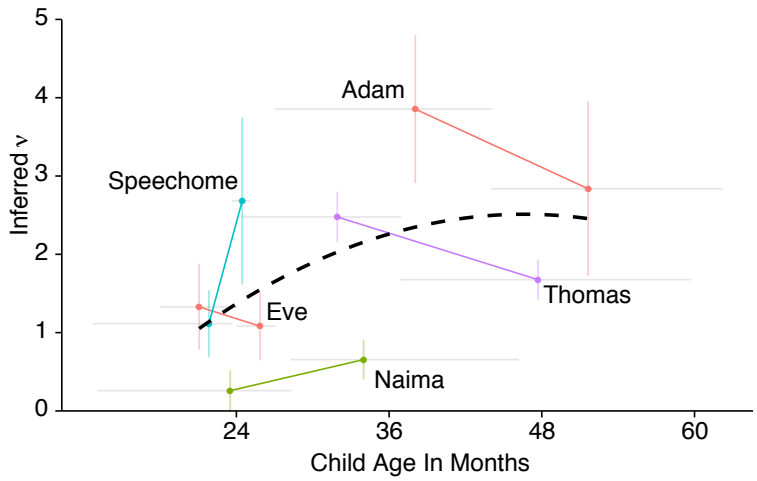


Figure 6. : The inferred developmental trajectory for determiner productivity, ν, across children reaching the convergence criterion after imitations and repetitions are filtered out ($n = 5$). Each line shows a two-point productivity trajectory for a single child, plotted by age in months. Marker size corresponds to the number of child tokens used for each child. Gray horizontal lines indicate the temporal extent of the tokens used to parameterize the model at each point; vertical lines indicate the SD of the posterior. The best fitting quadratic trend is shown as a dashed black line.

References

Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: If, when and why. *Cognitive Psychology*, *6*, 380-420.

Brown, R. (1973). *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language & Speech*, *49*(2), 137–173.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP* (p. 517-520).

Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday experience of young american children*. Brookes Publishing Company.

Kuczaj, S. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, *16*, 589–600.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, *20*(3), 481-508.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates.

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R., & Pennebaker, J. (2007). Are women really more talkative than men? *Science*, *317*, 82.

Pine, J., & Lieven, E. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, *18*(2), 123–138.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.*

Roy, B., Frank, M., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663-12668.

Roy, B., Frank, M., & Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. In *Proceedings of the 31st annual meeting of the cognitive science society.*

Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent–child discourse. In *Children's Language* (Vol. 4). Lawrence Erlbaum Associates.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of childes transcripts. *Journal of Child Language*, *37*(3), 705–729.

Suppes, P. (1974). The semantics of children's language. *American Psychologist*, *29*, 103-114.

Theakston, A., Lieven, E., Pine, J., & Rowland, C. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, *28*, 127–152.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL* (p. 252-259).

Vosoughi, S., & Roy, D. (2012). An automatic child-directed speech detector for the study of child language development. In *Proceedings of Interspeech.*

Yang, C. (2013). Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*.